Classification and Prediction of Survival in Hepatocellular Carcinoma by Gene Expression Profiling

Ju-Seog Lee,[1] In-Sun Chu,[1] Jeonghoon Heo,[1] Diego F. Calvisi,[1] Zongtang Sun,[2] Tania Roskams,[3] Anne Durnez,[3] Anthony J. Demetris,[4] and Snorri S. Thorgeirsson[1]

We analyzed global gene expression patterns of 91 human hepatocellular carcinomas (HCCs) to define the molecular characteristics of the tumors and to test the prognostic value of the expression profiles. Unsupervised classification methods revealed two distinctive subclasses of HCC that are highly associated with patient survival. This association was validated via 5 independent supervised learning methods. We also identified the genes most strongly associated with survival by using the Cox proportional hazards survival analysis. This approach identified a limited number of genes that accurately predicted the length of survival and provides new molecular insight into the pathogenesis of HCC. Tumors from the low survival subclass have strong cell proliferation and antiapoptosis gene expression signatures. In addition, the low survival subclass displayed higher expression of genes involved in ubiquitination and histone modification, suggesting an etiological involvement of these processes in accelerating the progression of HCC. In conclusion, the biological differences identified in the HCC subclasses should provide an attractive source for the development of therapeutic targets (*e.g.*, HIF1a) for selective treatment of HCC patients. *Supplementary material for this article can be found on the HEPATOLOGY Web site (http://interscience.wiley.com/jpages/0270-9139/suppmat/index.html).* (HEPATOLOGY 2004;40:667–676.)

Hepatocellular carcinoma (HCC) is the fifth most common cancer in the world, accounting for an estimated 500,000 deaths annually.[1] Although HCC is prevalent in Southeast Asia and sub-Sahara Africa, the incidence of HCC has doubled in the United States over the past 25 years, and incidence and mortality rates are likely to double over the next 10–20 years.[2] Although much is known about both the cellular changes that lead to HCC and the etiological agents responsible for the majority of HCC cases (hepatitis B virus, hepatitis C virus, alcohol), the molecular pathogenesis of HCC is not well understood.[3] Considerable efforts have been devoted to establishing a prognostic model for HCC by using clinical information and pathological classification to provide information at diagnosis on both survival and treatment options.[4–10] Although much progress has been made (reviewed by Llovet et al.[11]), many issues still remain unresolved. For example, a staging system that reliably separates patients with early HCC as well as intermediate to advanced HCC into homogeneous groups with respect to prognosis does not exist. This is particularly important because the natural course of early HCC is unknown and the natural progression of intermediate and advanced HCC are known to be quite heterogeneous.[12] It therefore appears axiomatic that improving the classification of HCC patients into groups with homogeneous prognosis would at least improve the application of currently available treatment modalities and at best provide new treatment strategies.

Recently, microarray technologies have been successfully used to predict clinical outcome and survival as well as classify different types of cancer.[13–15] These microarray

Abbreviations: HCC, hepatocellular carcinoma; ST, surrounding tissue; AFP, alpha fetoprotein; HA genes, genes with high expression in subclass A tumors; HB genes, genes with high expression in subclass B tumors.

From the [1]Laboratory of Experimental Carcinogenesis, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD; the [2]National Laboratory of Molecular Oncology, Cancer Institute, Chinese Academy of Medical Sciences, Beijing, China; the [3]Departments of Morphology and Molecular Pathology and Hepatology, University of Leuven, Leuven, Belgium; and the [4]Thomas E. Starzl Transplant Institute, University of Pittsburgh Medical Center, Pittsburgh, PA.

Received April 8, 2004; accepted May 5, 2004.

Address reprint requests to: Snorri S. Thorgeirsson, Laboratory of Experimental Carcinogenesis, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892-4255. E-mail: snorri_thorgeirsson@nih.gov; fax: (301) 496-0734.

Published online in Wiley InterScience (www.interscience.wiley.com).
DOI 10.1002/hep.20375

technologies have also been applied in many studies to define global gene expression patterns in primary human HCC as well as HCC-derived cell lines[16] in an attempt to gain insight into the mechanisms of hepatocarcinogenesis. These studies have identified subgroups of HCC that differ according to etiological factors,[17] mutations of tumor suppressor genes,[18] rate of recurrence,[19] and intrahepatic metastasis,[20] as well as novel molecular markers for HCC diagnosis.[21] However, most of these studies identified genes that are associated with limited aspects of tumor pathogenesis, and thus failed to create molecular prognostic indices that could be applied to the HCC patient population in general.

In the present study, we investigated the possibility that variations in gene expression in HCC obtained at diagnosis would permit the identification of distinct subclasses of HCC patients with different prognoses. The results revealed two subclasses of HCC patients characterized by significant differences in the length of survival. We also identified expression profiles of a limited number of genes that accurately predicted the length of survival. Our data indicate that it is possible to use gene expression patterns to accurately predict the clinical outcome of HCC at the time of diagnosis.

Patients and Methods

Complementary DNA Microarrays. The Human Array-Ready Oligo Set (Version 2.0) containing 70-mer probes of 21,329 genes was obtained from Qiagen, Inc. (Valencia, CA). Oligo microarrays were produced at the Advanced Technology Center at the National Cancer Institute.

Human Tissue Samples and Preparation of RNA. Surgically removed normal livers ($n = 18$) from patients with liver metastasis from colon cancers or from traffic accident patients were retrieved from the tissue bank of the Thomas E. Starzl Transplant Institute at the University of Pittsburgh Medical Center. One disease-free donor liver unsuitable for transplantation was also used. Total RNAs from the 19 normal livers were pooled and used as a reference for all microarray experiments. Ninety-one HCC tissues and 60 matched nontumor surrounding liver tissues were obtained from 90 patients undergoing partial hepatectomy as treatment for HCC. Tumor specimens originated from China and Belgium. Tissue banking was approved by the Institutional Review Board of all institutions. Total RNAs were isolated using the CsCl density gradient centrifugation method.[22]

Microarray Experiments and Data Analysis. Twenty micrograms of total RNA from tissues were used to derive fluorescently (Cy5 or Cy3) labeled complementary DNA. A reference complementary DNA was generated using total RNA from 19 normal livers. At least two hybridizations were performed for each tissue sample using a dye-swap strategy to eliminate labeling bias of the fluorescent intensity measurement. A detailed procedure for microarray experimentation and data analysis is available in a supplementary note on the HEPATOLOGY Web site (http://interscience.wiley.com/jpages/0270-9139/suppmat/index.html).

Supplementary Data. Supplementary notes, figures, and tables can be accessed on the HEPATOLOGY Web site (http://interscience.wiley.com/jpages/0270-9139/suppmat/index.html).

Results

We characterized gene expression profiles in 91 human primary HCC and 60 matched nontumor surrounding tissues (STs) using DNA microarrays. A hierarchical clustering analysis based on Pearson correlation coefficients was applied to all tissues on the basis of similarity in the expression pattern over all genes (Fig. 1A). As expected, it yielded two major clusters, one representing HCC tumors, and the other representing nontumor STs, with a few exceptions. Thus, the molecular configuration of HCC can be readily distinguished from nontumor STs, as has already been observed.[18]

Two Distinct Subclasses of HCC Revealed via Hierarchical Clustering of Gene Expression Patterns are Highly Associated With Survival of Patients. Next, we attempted to identify subclasses of HCC solely on the basis of gene expression patterns. Genes with an expression ratio that has at least a twofold difference relative to the reference in at least 9 tumors were selected for hierarchical analysis (4,187 gene features). Analysis of the clustered data with the HCC revealed 2 distinctive subtypes of gene expression patterns among 91 cases of HCC (Fig. 1B), suggesting a degree of heterogeneity among HCC gene expression profiles. Members of the 2 clusters also resided in compact and easily separable three-dimensional space when viewed by a three-dimensional multidimensional scaling plot based on their overall similarity of expression patterns (Supplementary Fig. 2), indicating that the 2 subclasses identified with hierarchical clustering are not due to artifacts from data processing. Having identified the 2 distinctive subclasses of HCC, we examined the association of clusters with clinical data. The two clusters showed weak associations with serum alpha fetoprotein (AFP) levels and Edmonson tumor grades. Cluster A contained a higher percentage of AFP+ (>300 ng/mL) patients (62.5%) and Edmonson grade III tumors (77%), while 42% and 50% of cluster B was AFP+ and grade III, respectively (Table 1). Significant association with the clusters was only detectable in patient survival. The over-

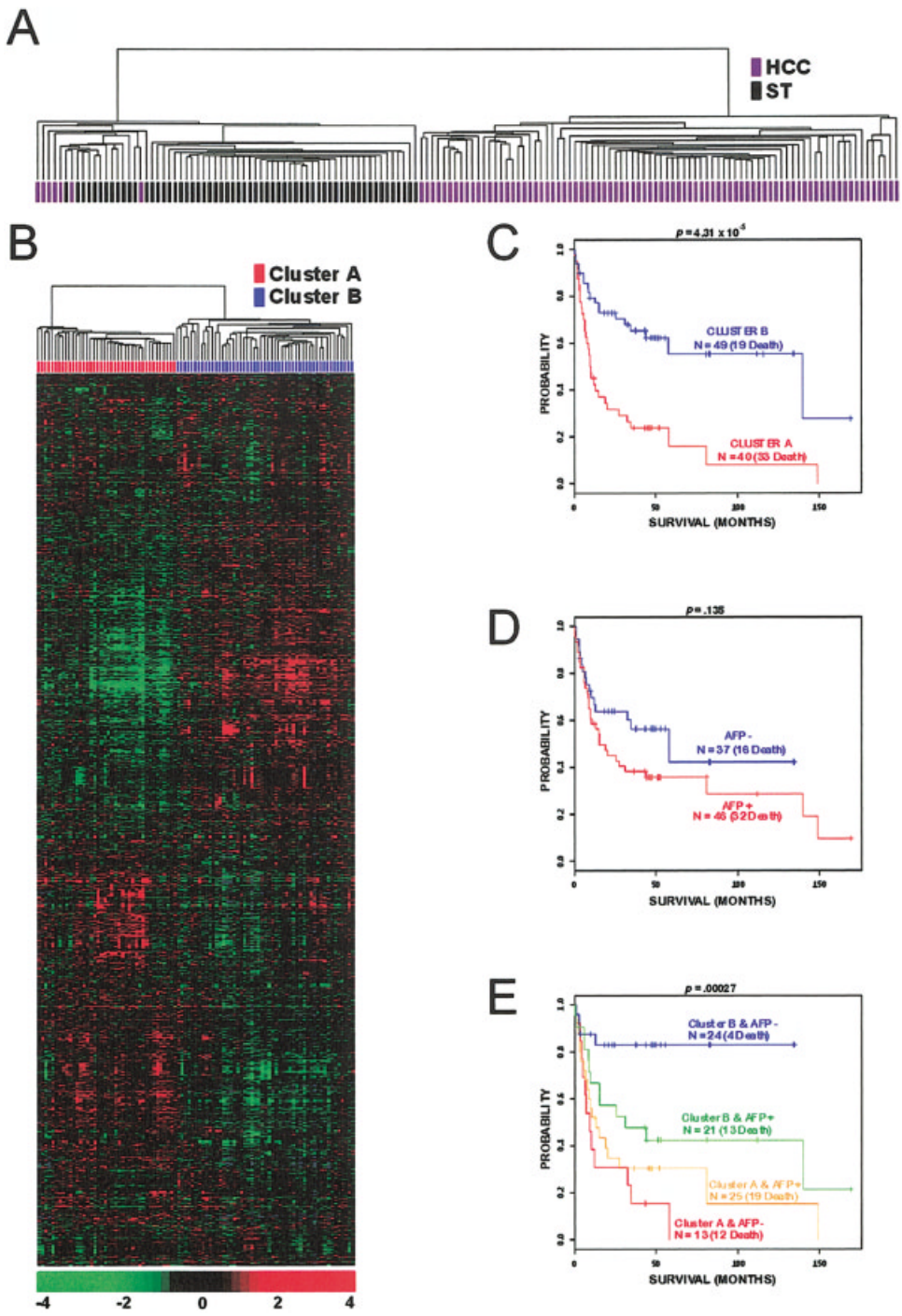

Fig. 1. Hierarchical clustering analysis. (A) Unsupervised hierarchical clustering of 91 HCC tumors and 60 matched surrounding nontumor liver tissues separated the tissues into two main groups: HCC tumors and STs (see Supplementary Fig. 1 for details). (B) Hierarchical clustering of 91 HCC tumors only. Genes with an expression ratio that had at least a twofold difference relative to reference in at least 9 tissues were selected for hierarchical analysis (4,187 gene features). The data are presented in matrix format in which rows represent the individual gene and columns represent each tissue. Each cell in the matrix represents the expression level of a gene feature in an individual tissue. The red and green color in cells reflect high and low expression levels, respectively, as indicated in the scale bar (log2 transformed scale). (C) Kaplan-Meier plot of overall survival of HCC patients grouped on the basis of gene expression profiling. One patient (HCC16) was excluded from the data set due to death from septic shock after surgery. (D) Kaplan-Meier plot of overall survival of HCC patients grouped on the basis of serum AFP levels (>300 ng/mL). (E) Kaplan-Meier plot of overall survival of HCC patients grouped on the basis of both gene expression profiling and AFP levels in serum. HCC, hepatocellular carcinoma; ST, surrounding tissue; AFP, alpha fetoprotein.

all survival time in cluster A (30.3 ± 8.02 months) was shorter than cluster B (83.7 ± 10.3 months). As expected, a Kaplan-Meier survival curve and a log-rank test indicated poorer survival in cluster A patients ($P < 1.0 \times 10^{-4}$) when compared with cluster B (Fig. 1C). Thus, the molecular differences between these 2 subclasses of HCC were associated with a remarkable difference in the clinical outcome of these patients.

It has been widely accepted that serum AFP levels are significantly related to the survival of HCC patients; higher levels of serum AFP indicate poorer survival.[5,6,10] Among many clinical indicators of the HCC patients, serum AFP levels (>300 ng/mL or less) only showed association with survival with marginal significance ($P = .13$) in our patient cohort (Fig. 1D). We determined whether or not our molecular classification of HCC could enhance the prognostic value of this clinical indicator previously used for prediction of survival. While the 2 subclasses of AFP+ patients showed a marginal difference in overall survival, AFP− patients in cluster A had a severely

Table 1. Clinical and Pathological Features of HCC Patients

Variable	Cluster A	Cluster B	Total
No. of patients	40	50	90
Male	33	38	71
Female	7	12	19
Age			
Mean	51.2	54	52.7
SD	11.2	13	12.2
AFP (>300 ng/mL)			
+	25	21	46
−	13	25	38
NA	2	4	6
Etiology			
HBV	29	23	52
HCV	2	5	7
HBV/HCV	–	3	3
Alcohol	–	5	5
HBV/Alcohol	–	1	1
HCV/Alcohol	–	1	1
Hereditary hemochromatosis	1	2	3
Wilson's disease	1		1
NA	7	10	17
Edmonson grade			
II	8	24	32
III	31	25	56
IV	1	1	2
Cirrhosis			
+	21	24	45
−	19	26	45
Death	33	20	53
Survival (months)*			
Mean	30.5	83.7	64.6
SE	8.02	10.3	8.53

Abbreviations: HBV, hepatitis B virus; HCV, hepatitis C virus; NA, not applicable.
*Mean and SE of survival time were estimated as described in supplementary notes. See details in Supplementary Table 3.

diminished overall survival (Supplementary Fig. 3). Moreover, when patients were subdivided into 4 groups based on serum AFP levels and gene expression clusters, AFP− patients in cluster A showed the worst overall survival among all patients (Fig. 1E).

Prediction of Survival With Gene Expression Profiles. We applied 5 different statistical methods to determine whether or not gene expression patterns could be used to predict survival: linear discriminator analysis, support vector machines, nearest centroid, nearest neighbor, and compound covariate predictor. Before the analysis, 2 tumor samples, HCC89-2 and HCC16, were excluded from the data set because patient HCC16 died of septic shock after surgery and samples from 2 separate tumors were obtained from patient HCC89. In the absence of a totally independent data set, we attempted to assess the validation of our results and reproducibility of the test by randomly dividing the HCC into 2 equal groups: the training set ($n = 45$), which was used to develop the HCC classifiers, and the validation set ($n = 44$), which was used to evaluate the test. Briefly, we started to identify the most differentially expressed genes between 2 clusters in the training set. These genes were combined to form a series of classifiers that estimate the probability that a particular HCC belongs to cluster A or B. The number of genes in the classifiers was optimized to minimize misclassification errors during the "leave one out" cross-validation of the tumors in the training set. When applied to the validation set, all 5 models successfully separated poorer survival patients (cluster A) from longer survival patients (cluster B). All Kaplan-Meier survival curves and log-rank tests in the validation set showed significant differences between subclass A and B that were independently predicted using the 5 classifier models (Fig. 2A–G). Moreover, when we examined the predicted subclass memberships of the tumors, only a few discrepancies were observed (Fig. 2H). These results demonstrated not only strong association of gene expression patterns with the survival of the patients but also a robust reproducibility of these gene expression–based predictors.

Survival Genes. Because the most striking feature of the unsupervised analysis of the expression profiles was the strong association with survival, we decided to apply supervised analysis of the genes whose expression is most strongly associated with length of survival. The univariate Cox proportional hazards model was used to assess the association of the gene expression with the survival. Expression of 442 features representing 406 unique genes (Supplementary Table 1) was highly correlated with length of survival with strong statistical significance ($P < .001$). The outcome of hierarchical cluster analysis of the HCC with the 406 survival genes was highly similar to the previous analysis with all the genes (Fig. 3A). With few exceptions, cluster memberships of each tumor remained the same in the 2 hierarchical cluster dendrograms, highlighting again the robustness of the predicted HCC subclasses and their strong association with length of survival. We noted that survival genes were almost equally divided into two groups, those whose expression is higher in subclass A tumors (HA genes) and those whose expression is higher in subclass B tumors (HB genes). When we categorized the survival genes according to the Gene Ontology, the biggest difference between HA genes and HB genes was observed in genes associated with cell proliferation (Supplementary Table 2). Of the HA survival genes, 45% belonged to the cell growth and maintenance category, while only 19% of HB survival genes were in the same Gene Ontology category, strongly suggesting that the HCCs in subclass A grow faster than those in subclass B.

We next generated an averaged gene expression index from HB genes to examine their predictive power. Patients were then ranked according to the average gene expression level of tumors from the highest to the lowest (Fig. 3B) and divided into two equal 50th percentiles.

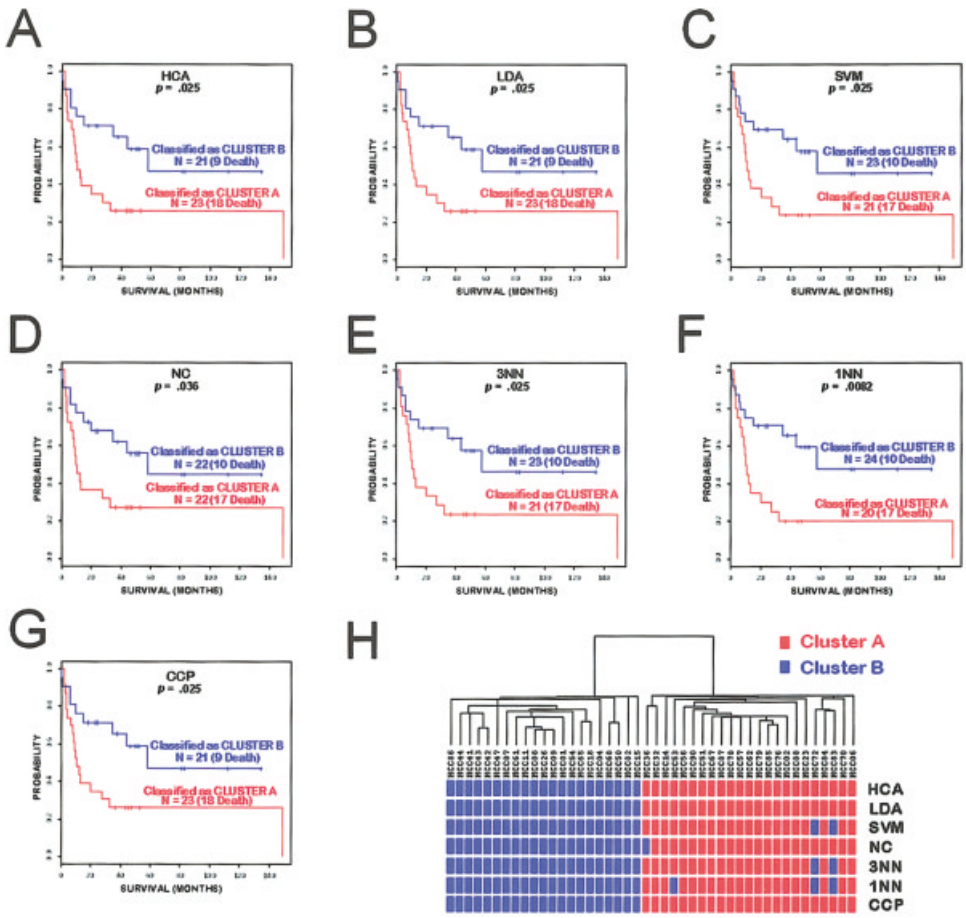

Fig. 2. Survival analysis of outcome of prediction in validation set. (A) Kaplan-Meier plot of overall survival of HCC patients in validation set classified by hierarchical clustering analysis. (B–G) Kaplan-Meier plots of overall survival of HCC patients in validation set classified by linear discriminator analysis, support vector machines, nearest centroid, 3 nearest neighbor, 1 nearest neighbor, and compound covariate prediction models, respectively. (H) Hierarchical clustering of 44 HCC tissues in a validation set. Columns represent each tissue and rows represent outcomes of various prediction models as indicated. Each cell represents memberships of tissues when particular prediction model was applied in a validation set. The red and blue color in cells represent clusters A and B, respectively. HCA, hierarchical clustering analysis; LDA, linear discriminator analysis; SVM, support vector machines; NC, nearest centroid; NN, nearest neighbor; CCP, compound covariate prediction.

Kaplan-Meier plots and log-rank tests of overall patient survival in the 2 divided groups revealed striking differences with strong statistical significance ($P < 1.0 \times 10^{-4}$) (data not shown). Likewise, the average gene expression index from HA genes produced similar results (Fig. 3C) with comparable statistical significance ($P < 1.0 \times 10^{-5}$). Most of the patients in cluster A and B were perhaps not surprisingly well separated from each other in both 50th percentile segmentations (Fig. 3B and 3C). Taken together with the previous 2 independent clustering analyses and the cross-validation test of training and validation data sets, these results further support the notion that a distinct gene expression pattern predicts survival characteristics of the 2 subclasses of the HCC patients.

Next, we employed a knowledge-based annotation of the survival genes based on a public database search, because the Gene Ontology Consortium term annotation of genes was not sufficient to provide insight into the underlying biological differences between the 2 subclasses of HCC. The survival genes fell within several biological groups (Table 2). The cell proliferation group was the best predictor of an unfavorable outcome of the disease, which is consistent with previous analyses in human lymphomas.[23] Expression of typical cell proliferation markers such as *PNCA* and cell cycle regulators such as *CDK4*, *CCNB1*, *CCNA2*, and *CKS2* was greater in subclass A than subclass B. Not surprisingly, many genes that are expressed more in subclass A are antiapoptotic. Recent studies have identified *PTMA*/*ProT* as an inhibitor of apoptosome formation, the essential step for the final activation of the caspase-dependent cascade in the apoptotic pathway,[24] and have identified *SET* as an inhibitor of the Granzyme A–induced caspase-independent pathway.[25] *SET* is also a subunit of the inhibitor of acetyltransferases complex that regulates histone modification and gene expression.[26] Significantly, *PTMA* has recently been shown also to be part of the inhibitor of acetyltransferases complex,[27] suggesting their multiple roles in hepatocarcinogenesis. Genes involved in prothrombin activation were expressed less in subclass A, indicating impairment of liver function in this subclass. Many of the genes with lower expression in subclass A were liver-specific (data not shown), which is consistent with the previous observation that poorly differentiated HCC tumors have less favorable clinical outcomes.[28] Higher expression of genes involved in ubiquitination and sumoylation indicated that accelerated cell proliferation in the poorer survival group might be due to selective degradation of critical proteins, including cell cycle inhibitors. Concomitant overexpression of the histone H4 family with *HRMT1L2* (H4-specific

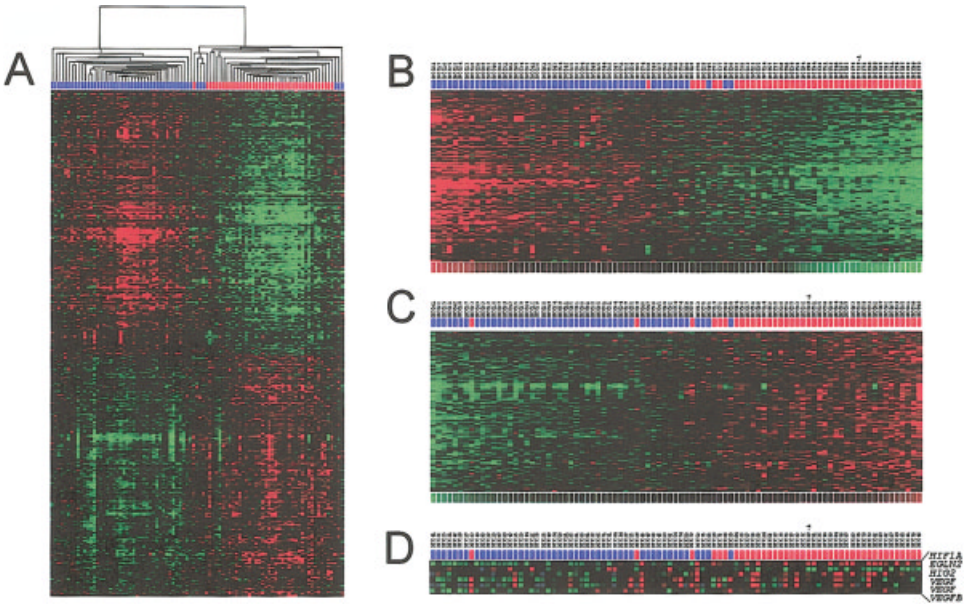

Fig. 3. Gene expression patterns of 406 survival genes. (A) Hierarchical clustering of 89 HCC tumors with survival genes separated the tissues into 2 main groups. The data are presented in as described in Fig. 1. The red and blue color cells at the bottom of dendrogram represent memberships of cluster A and B, respectively, from Fig. 1B. (B) Relative expression of 213 HB genes that were more expressed in cluster B HCC tissues. HCC tissues were ordered according to average expression level of 213 genes as indicated at the bottom of colored heat map. (C) Relative expression of 193 HA genes that were more expressed in cluster A HCC tissues. HCC tissues were ordered according to average expression level of 193 genes as indicated at the bottom of colored heat map. (D) Relative expression of *HIF1a*, *ENGL2*, and downstream target genes of *HIF1a*. HCC tissues were ordered according to average expression level of 193 genes as indicated.

methyltransferase) in the poorer survival group of HCC may indicate unidentified roles of the histone H4 family and their modification in tumor development. Expression of *HIF1a*, the master regulator of hypoxia induced gene expression,[29] was enhanced in subclass A, while expression of *ENGL2*, a negative regulator of *HIF1a* by prolyl hydroxylation,[30] was reduced (Fig. 3D). Both these changes dramatically enhance HIF1a activity in tumor cells, which in turn provide a favorable environment for tumor growth.

Predicted biological features of each subgroup of HCC based on gene expression patterns were further validated using independent methods as described in the supplementary notes.

Three Distinctive Gene Expression Patterns in HCC and ST. To gain additional insight into the biological differences between the 2 subclasses of HCC, we generated 2 different gene lists by applying significance analysis of microarrays.[31] Gene list X represents the top 500 genes that were differentially expressed between ST and all HCC tissues. Gene list Y represents the top 500 genes that were differentially expressed between HCCs in A and B clusters (Fig. 4A and 4B). When gene expression patterns of all tissues were compared together, 3 different patterns were observed: X not Y (330 genes), X and Y (170 genes), and Y but not X (330 genes). Genes in the X not Y category had uniform differences between all HCCs and STs regardless of subclass A or B, representing common alterations of gene expression in HCC. Enhanced expression of 26S proteasome subunits such as *PSMC4*, *PSME3*, *PSMD4*, *PSMD2*, and *PSMB4* indicated an enhanced activation of wide-ranging protein degradation inall HCCs. However, ubiquitination, a selective protein degradation pathway, was only active in subclass A HCC (see Table 2). Genes in the X and Y category display a subclass-specific gene expression pattern. Although gene expression was altered in all HCCs, a more pronounced alteration was observed in subclass A. Expression of the G1/S phase cell cycle regulator *CDK4* was highest in subclass A , moderately enhanced in subclass B, and lowest in ST, which agreed well with the more proliferative features of subclass A. Enhanced expression of *H2FAX*—a histone H2 variant involved in the chromosome double-strand breaks response[32]—in subclass A might reflect more chromosomal damage and/or instability in subclass A than in B. Additional reduction in the expression of liver-specific genes including the p450 family in the poorer survival group (subclass A) shows that reduced liver function is indeed a bad prognostic indicator for HCC patients.

Discussion

Prognostic modeling of patients with HCC at diagnosis that considers tumor stage, functional impairments of

Table 2. Summary of Selected Survival Genes

Gene	Hazard Ratio	*P* (Wald test)	*P* (Likelihood Ratio test)	*P* (*t* test, A vs. B)	Unigene	Description
Prothrombin activation						
F10	0.718	.00093	.00136	3.63E-07	Hs.47913	Coagulation factor X
F12	0.713	.000064	8.39E-05	5.32E-15	Hs.1321	Coagulation factor XII (Hageman factor)
KNG	0.709	.000096	.000204	2.35E-15	Hs.77741	Klninogen
SERPINC1	0.774	.00026	.000303	3.49E-20	Hs.75599	Serine (or cysteine) proteinase inhibitor, clade C1
SERPING1	0.586	.00022	.000273	2.26E-12	Hs.151242	Serine (or cysteine) proteinase inhibitor, clade G1
Ubiquitination and sumoylation						
UBE2D1	2.4	.00034	.000303	3.15E-12	Hs.129683	Ubiquitin-conjugating enzyme E2D 1
USP1	2.22	.00061	.000951	3.07E-06	Hs.35086	Ubiquitin specific protease 1
HSPC150	1.76	.00074	.00105	8.31E-08	Hs.5199	Similar to ubiquitin-conjugating enzyme
UBA2	2.43	7.40E-07	1.17E-06	9.28E-12	Hs.4311	SUMO-1 activating enzyme subunit 2
RBX1	2.32	.00017	.00014	3.39E-07	Hs.279919	Ring-box 1
RWDD1	2.34	5.30E-06	5.52E-05	.002542	Hs.22679	RWD domain containing 1
SMT3H2	3.28	.0003	.000255	1.77E-08	Hs.180139	SUMO-2
Histones						
HIST1H4A	1.97	.000049	3.98E-05	4.51E-11	Hs.248178	Histone 1, H4a
HIST1H4C	1.69	.00051	.000306	1.91E-10	Hs.46423	Histone 1, H4c
HIST2H4	2.04	.00013	.000141	6.44E-11	Hs.55466	Histone 2, H4
HRMT1L2	2.31	.00065	.00087	3.38E-14	Hs.20521	HMT1 hnRNP methyltransferase-like 2
CRFG	3.77	.000035	2.74E-05	6.51E-12	Hs.215766	G protein-binding protein CRFG
HDAC2	2.64	.000015	2.42E-05	4.79E-12	Hs.3352	Histone deacetylase 2
SLBP	2.93	.00039	.000586	.001256	Hs.75257	Stem-loop (histone) binding protein
Apoptosis						
PTMA	3.75	.000038	3.96E-05	7.89E-08	Hs.250655	Prothymosin, alpha
SET	2.2	.0007	.0011	2.65E-07	Hs.145279	SET translocation
YWHAB	3.48	.00046	.000408	5.25E-06	Hs.182238	14-3-3 beta polypeptide
YWHAH	2.44	.00063	.000347	1.18E-08	Hs.349530	14-3-3 eta polypeptide
YWHAQ	2.31	.00085	.000656	2.81E-11	Hs.74405	14-3-3 theta polypeptide
NALP2	2.91	.00035	.000164	1.77E-11	Hs.6844	Neuronal apoptosis inhibitor protein 2
PDCD5	3.04	.000011	1.41E-05	6.25E-06	Hs.166468	Programmed cell death 5
P8	0.525	.00073	.000704	.000309	Hs.424279	p8 Protein (candidate of metastasis 1)
IER3	1.74	.00038	.000295	8.43E-11	Hs.76095	Immediate early response 3
Cell cycle regulation and cell proliferation						
PCNA	1.92	.00022	.000301	3.57E-06	Hs.78996	Proliferating cell nuclear antigen
CDK4	2.09	.00085	.000836	3.78E-12	Hs.95577	Cyclin-dependent kinase 4
TOPBP1	4.24	.0001	6.85E-05	3.57E-08	Hs.91417	Topoisomerase (DNA) II binding protein
CGR11	0.422	.00038	.00029	1.57E-06	Hs.159525	Cell growth regulatory with EF-hand domain
BCAT1	1.58	.000081	.000254	4.35E-08	Hs.317432	Branched chain aminotransferase 1, cytosolic
CCNB1	2.19	.000058	7.66E-05	6.66E-07	Hs.23960	Cyclin B1
CKS2	1.77	.00011	.000115	9.41E-10	Hs.83758	CDC28 protein kinase regulatory subunit 2
DLG7	2.69	.000017	6.12E-05	4.3E-07	Hs.77695	Discs, large homolog 7 (Drosophila)
NAP1L1	1.98	.000074	.000111	6.42E-10	Hs.302649	Nucleosome assembly protein 1-like 1
CCNA2	1.97	.00015	.000335	8.11E-10	Hs.85137	Cyclin A2
MAPRE1	2.68	.000014	9.43E-06	3.29E-13	Hs.234279	Microtubule-associated protein, RP/EB family, member 1
TTK	1.66	.00095	.00128	1.08E-10	Hs.169840	TTK protein kinase
BUB3	2.69	.00083	.00104	8.13E-09	Hs.40323	Budding uninhibited by benzimidazoles 3 homolog
CENPF	1.71	.00047	.000593	4.53E-07	Hs.77204	Centromere protein F, 350/400ka
KNTC1	2.35	.00012	.000204	4.47E-05	Hs.333355	Kinetochore associated 1
NPM1	2.14	.00025	8.52E-05	3.49E-06	Hs.355719	Nucleophosmin
MCM2	2.03	.00099	.00119	8.1E-08	Hs.57101	Minichromosome maintenance deficient 2
MCM6	1.88	.00094	.000852	2.3E-10	Hs.155462	Minichromosome maintenance deficient 6
MCM7	2.41	.00047	.000434	2.6E-09	Hs.77152	Minichromosome maintenance deficient 7
Regulation of HIF1a						
HIF1A	1.69	.00022	.000298	1.19E-07	Hs.197540	Hypoxia-inducible factor 1, alpha
EGLN2	0.461	.00018	.000256	1.1E-08	Hs.324277	egl nine homolog 2

the liver, and the general condition of the patient can provide valuable information and indicate therapy.[4,10] However, increased surveillance and advances in image technology have afforded earlier diagnosis of HCC. This development presents a challenge with respect to prognostic modeling of HCC, because the natural history of early HCC is unknown.[12] In addition, intermediate and advanced HCC are quite heterogeneous,[33] even though

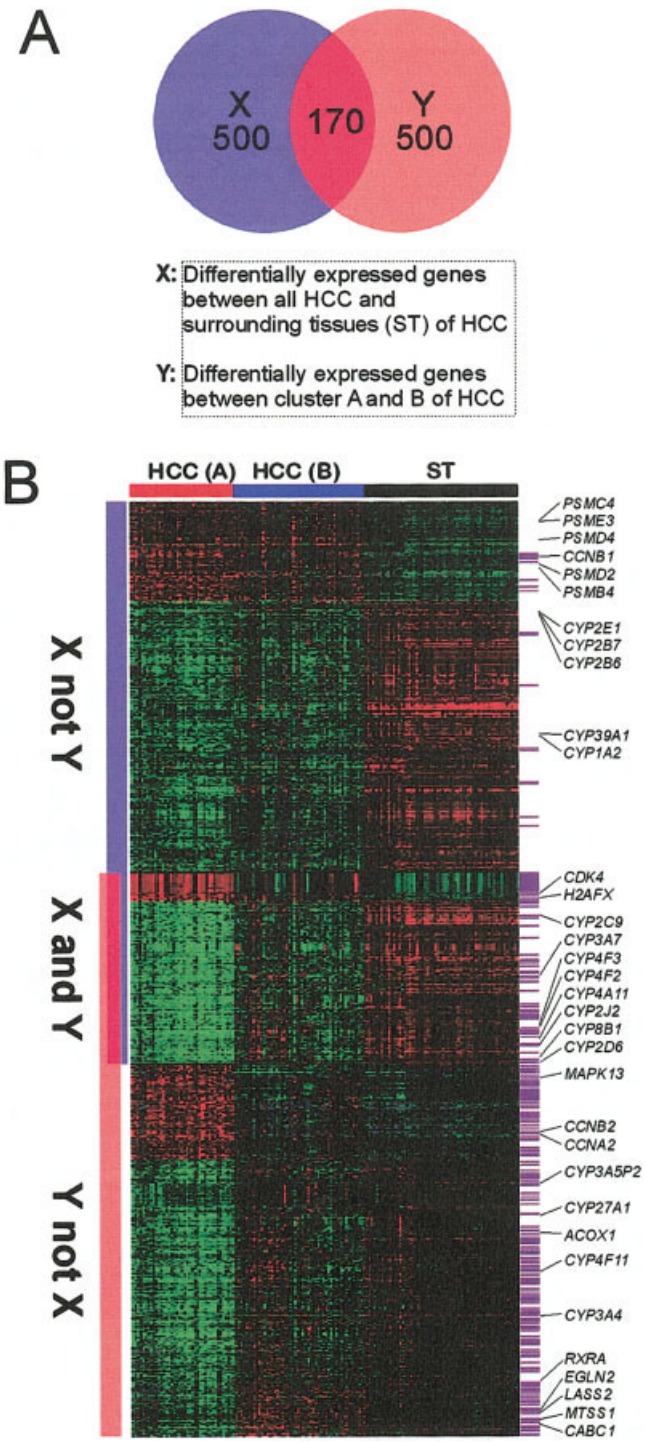

Fig. 4. Distinctive gene expressions of HCC and ST. (A) Venn diagram of genes selected via significance analysis of microarrays. X represents genes differentially expressed between surrounding tissues and all HCC tissues. Y represents genes differentially expressed between cluster A HCC and cluster B HCC tissues. One hundred seventy genes were shared in 2 different gene lists. (B) Purple and red bars at the left side of the heat map represent X and Y genes, respectively. Pink bars at the right side of heat map represent survival genes. Colored bars at the top of heat map represent tissues as indicated. HCC, hepatocellular carcinoma; ST, surrounding tissue.

the natural history and prognostic factors are well defined.[12] Therefore, it is necessary to establish robust methods capable of evaluating the prognosis of patients diagnosed at the early, intermediate, and late stages of HCC. As a first step in the development of a molecular prognostic evaluation, we have used gene expression profiling technology and unsupervised and supervised learning methods to successfully predict survival of HCC patients.

We applied three independent but complementary approaches for data analysis to uncover subclasses of HCC and the underlying biological differences between the subclasses. First, unsupervised classification methods based solely on gene expression patterns were applied. Hierarchical clustering of the data as well as multidimensional scaling revealed two subclasses of HCC strongly associated with the length of patients' survival. The differences in gene expression are quite robust as illustrated by the fact that the poorer survival group (subclass A) was successfully separated from the better survival group (subclass B) in the training data set as well as in the validation data set when 5 different statistical methods for prediction were applied (see Fig. 2). The presence of two extreme subgroups in AFP− patients was unexpected and probably accounts for the insufficient predictive power when AFP was used as a sole prognostic indicator.[4,10]

Second, a univariate regression model was used to identify individual genes whose expression is highly correlated with length of survival. Application of survival genes for subclass prediction was highly accurate, as illustrated by the fact that averaged gene expression indices were sufficient to segregate the 2 subclasses even without the use of sophisticated prediction models. Also, information obtained from knowledge-based annotation of the 406 survival genes provided insight into the underlying biological differences between the 2 subclasses of HCC. Although quantitative measurement of cell proliferation and apoptotic rates in both subclasses strongly support the long-established notion that imbalance between cell proliferation and cell death is the primary hallmark of tumors[34] and provided the best quantitative separation of the 2 subclasses, many other issues were also highlighted.

The ubiquitin system is often deregulated in cancers.[35] In HCC, the degree of ubiquitination is highly correlated with cell proliferation and survival of patients and has also been proposed as a possible predictive marker for recurrence of human HCC.[36] In addition, PSMD10/Gankyrin, a subunit of the 26S proteasome that accelerates the degradation of retinoblastoma, is overexpressed in HCC.[37] Also, enhanced activation of ubiquitin-dependent protein degradation may account for deregulation of cell cycle control and faster cell prolifera-

tion in the poor survival group (subclass A). Therefore, deregulated components in ubiquitin-mediated protein degradation may provide attractive therapeutic targets for novel HCC treatment modalities.

The third approach involved the analysis of overall similarity and dissimilarity of gene expression between all the HCCs, subclasses A and B, and STs (see Fig. 4). Although the 2 subclasses of HCC may be viewed as distinctive biological entities, they still share significant overall similarity of gene expression when compared with ST. This may indicate that subclass A HCC accumulated additional oncogenic alterations of gene expression on top of a common HCC gene expression signature, thereby providing a more favorable environment for tumor cell growth. However, we cannot rule out the possibility that different mechanisms may contribute to the development of subclass A and B following exposure to different etiological factors, and the gene expression signature may reflect that etiological "footprint." This scenario is unlikely, however, because the great majority of our HCC cases were associated with hepatitis B virus. Alternatively, the cell of origin of a tumor can be important in determining the clinical outcome, as shown for diffuse large B cell lymphoma.[14] It is therefore possible that the 2 subclasses of HCC might represent different cellular origins (*i.e.*, hepatic stem cells vs. hepatocytes) of the tumors.

Comparative analysis of our data and earlier studies demonstrated good concordance of the data despite differences in patient populations and technology platforms (see supplementary notes). It strongly supports the generality of our findings that the subclasses of HCC might represent distinct disease entities. Also, the observation that genes associated with early recurrence and intrahepatic metastasis of HCC[19,20] did not discriminate between the subclass A and B suggests that the information (at least from a gene expression standpoint) embedded in these important processes is not sufficient to predict survival. It is therefore likely that the additional information provided by the survival genes (only 2 of the genes associated with intrahepatic metastasis were among these) is needed for effectively predicting survival. This is of considerable importance, because in a recent study on survival of HCC patients it was demonstrated that HCC was the prime cause of death in patients with compensated cirrhosis.[38] However, considerable molecular heterogeneity still exists within each HCC subclass, as evidenced by quantitative differences in survival gene expression (see Fig. 3B and 3C) and the small fraction of patients that are frequently misclassified in the prediction models. It is therefore probable that more subclasses of HCC might emerge when gene expression data from more HCC patients become available.

The severity of HCC and the lack of good diagnostic markers and treatment strategies have rendered the disease a major challenge. Systematic analysis of gene expression patterns provides an insight into the biology and pathogenesis of HCC. Our results indicate that HCC prognosis can be readily predicted from the gene expression profiles of the primary tumors. Because the microarray-based measurement of gene expression reflects the abundance of expressed messenger RNA and proteins in the HCC as confirmed by quantitative reverse-transcriptase polymerase chain reaction and immunohistochemical staining (Supplementary Figs. 6 and 7), a limited set of quantitative reverse-transcriptase polymerase chain reaction and/or immunohistochemical staining assays may be sufficient to predict the prognosis of patients at the time of diagnosis; however, a prospective study is needed to confirm this proposal. Nevertheless, the unique molecular characteristics of each subclass of HCC uncovered by a genome-wide survey of gene expression provide insight into the tumor biology of HCC and offer the opportunity for new therapeutic strategies. *SET* and *PTMA* are of particular interest for potential therapeutic targets because of their multitasking features. Even if a curative therapy for HCC patients cannot be offered at this stage, it may be possible to identify therapeutic targets that can slow the course of disease progression. For example, small molecules that inhibit *PTMA* and HIF1a activities are already available[24,39] and may provide opportunities to alter the course of HCC progression in both subclass A and B.

References

1. Parkin DM, Bray F, Ferlay J, Pisani P. Estimating the world cancer burden: Globocan 2000. Int J Cancer 2001;94:153–156.
2. El Serag HB, Mason AC. Rising incidence of hepatocellular carcinoma in the United States. N Engl J Med 1999;340:745–750.
3. Thorgeirsson SS, Grisham JW. Molecular pathogenesis of human hepatocellular carcinoma. Nat Genet 2002;31:339–346.
4. Bruix J, Llovet JM. HCC surveillance: who is the target population? HEPATOLOGY 2003;37:507–509.
5. Calvet X, Bruix J, Gines P, Bru C, Sole M, Vilana R, et al. Prognostic factors of hepatocellular carcinoma in the west: a multivariate analysis in 206 patients. HEPATOLOGY 1990;12:753–760.
6. Chevret S, Trinchet JC, Mathieu D, Rached AA, Beaugrand M, Chastang C. A new prognostic classification for predicting survival in patients with hepatocellular carcinoma. Groupe d'Etude et de Traitement du Carcinome Hepatocellulaire. J Hepatol 1999;31:133–141.
7. Okuda K, Ohtsuki T, Obata H, Tomimatsu M, Okazaki N, Hasegawa H, et al. Natural history of hepatocellular carcinoma and prognosis in relation to treatment. Study of 850 patients. Cancer 1985;56:918–928.
8. Pugh RN, Murray-Lyon IM, Dawson JL, Pietroni MC, Williams R. Transection of the oesophagus for bleeding oesophageal varices. Br J Surg 1973; 60:646–649.
9. Tan CK, Law NM, Ng HS, Machin D. Simple clinical prognostic model for hepatocellular carcinoma in developing countries and its validation. J Clin Oncol 2003;21:2294–2298.

10. CLIP Investigators. A new prognostic system for hepatocellular carcinoma: a retrospective study of 435 patients: the Cancer of the Liver Italian Program (CLIP) investigators. HEPATOLOGY 1998;28:751–755.
11. Llovet JM, Burroughs A, Bruix J. Hepatocellular carcinoma. Lancet 2003; 362:1907–1917.
12. Llovet JM, Bru C, Bruix J. Prognosis of hepatocellular carcinoma: the BCLC staging classification. Semin Liver Dis 1999;19:329–338.
13. van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415:530–536.
14. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000;403:503–511.
15. Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat Med 2002;8:816–824.
16. Lee JS, Thorgeirsson SS. Functional and genomic implications of global gene expression profiles in cell lines from human hepatocellular cancer. HEPATOLOGY 2002;35:1134–1143.
17. Okabe H, Satoh S, Kato T, Kitahara O, Yanagawa R, Yamaoka Y, et al. Genome-wide analysis of gene expression in human hepatocellular carcinomas using cDNA microarray: identification of genes involved in viral carcinogenesis and tumor progression. Cancer Res 2001;61:2129–2137.
18. Chen X, Cheung ST, So S, Fan ST, Barry C, Higgins J, et al. Gene expression patterns in human liver cancers. Mol Biol Cell 2002;13:1929–1939.
19. Iizuka N, Oka M, Yamada-Okabe H, Nishida M, Maeda Y, Mori N, et al. Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection. Lancet 2003;361: 923–929.
20. Ye QH, Qin LX, Forgues M, He P, Kim JW, Peng AC, et al. Predicting hepatitis B virus-positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning. Nat Med 2003;9: 416–423.
21. Smith MW, Yue ZN, Geiss GK, Sadovnikova NY, Carter VS, Boix L, et al. Identification of novel tumor markers in hepatitis C virus-associated hepatocellular carcinoma. Cancer Res 2003;63:859–864.
22. Sambrook J, Fritsch E, Maniatis T. Extration and Purification of RNA. Molecular Cloning. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press, 1989:7.19–7.22.
23. Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. N Engl J Med 2002;346:1937–1947.
24. Jiang X, Kim HE, Shu H, Zhao Y, Zhang H, Kofron J, et al. Distinctive roles of PHAP proteins and prothymosin-alpha in a death regulatory pathway. Science 2003;299:223–226.
25. Fan Z, Beresford PJ, Oh DY, Zhang D, Lieberman J. Tumor suppressor NM23-H1 is a granzyme A-activated DNase during CTL-mediated apoptosis, and the nucleosome assembly protein SET is its inhibitor. Cell 2003;112:659–672.
26. Seo SB, McNamara P, Heo S, Turner A, Lane WS, Chakravarti D. Regulation of histone acetylation and transcription by INHAT, a human cellular complex containing the set oncoprotein. Cell 2001;104:119–130.
27. Chakravarti D, Hong R. SET-ting the stage for life and death. Cell 2003; 112:589–591.
28. Mise K, Tashiro S, Yogita S, Wada D, Harada M, Fukuda Y, et al. Assessment of the biological malignancy of hepatocellular carcinoma: relationship to clinicopathological factors and prognosis. Clin Cancer Res 1998; 4:1475–1482.
29. Wang GL, Jiang BH, Rue EA, Semenza GL. Hypoxia-inducible factor 1 is a basic-helix-loop-helix-PAS heterodimer regulated by cellular O2 tension. Proc Natl Acad Sci U S A 1995;92:5510–5514.
30. Bruick RK, McKnight SL. A conserved family of prolyl-4-hydroxylases that modify HIF. Science 2001;294:1337–1340.
31. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 2001;98:5116–5121.
32. Rogakou EP, Boon C, Redon C, Bonner WM. Megabase chromatin domains involved in DNA double-strand breaks in vivo. J Cell Biol 1999; 146:905–916.
33. Llovet JM, Bustamante J, Castells A, Vilana R, Ayuso MC, Sala M, et al. Natural history of untreated nonsurgical hepatocellular carcinoma: rationale for the design and evaluation of therapeutic trials. HEPATOLOGY 1999; 29:62–67.
34. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell 2000;100:57–70.
35. Pagano M, Benmaamar R. When protein destruction runs amok, malignancy is on the loose. Cancer Cell 2003;4:251–256.
36. Shirahashi H, Sakaida I, Terai S, Hironaka K, Kusano N, Okita K. Ubiquitin is a possible new predictive marker for the recurrence of human hepatocellular carcinoma. Liver 2002;22:413–418.
37. Higashitsuji H, Itoh K, Nagao T, Dawson S, Nonoguchi K, Kido T, et al. Reduced stability of retinoblastoma protein by gankyrin, an oncogenic ankyrin-repeat protein overexpressed in hepatomas. Nat Med 2000;6:96–99.
38. Sangiovanni A, Del Ninno E, Fasani P, De Fazio C, Ronchi G, Romeo R, et al. Increased survival of cirrhotic patients with a hepatocellular carcinoma detected during surveillance. Gastroenterology 2004;126:1005–1014.
39. Semenza GL. Targeting HIF-1 for cancer therapy. Nat Rev Cancer 2003; 3:721–732.